Universidade do Minho
Escola de Ciências

Joana Filipa Costeira Paulo

Stability of Intrinsically Disordered Regions of two Transcription Factors by Molecular Dynamics

outubro de 2016

Joana Filipa Costeira Paulo **Stability of Intrinsically Disordered Regions of two Transcription Factors by Molecular Dynamics**

UMinho | 2016

Universidade do Minho
Escola de Ciências

Joana Filipa Costeira Paulo

Stability of Intrinsically Disordered Regions of two Transcription Factors by Molecular Dynamics

Dissertação de Mestrado
Mestrado em Biofísica e Bionanossistemas

Trabalho realizado sob orientação do
Professor Doutor Luís Silvino Alves Marques
e do
Professor Doutor Lennart Nilsson

outubro de 2016

Acknowledgements

I would like to start by thanking my supervisor Professor Lennart Nilsson. Thank you for giving me the opportunity of doing this project and accepted me in his group.

I have to thank Evdokiya Salamanova for teaching me practically everything I needed to know in order to be able to work independently, Professor Anthony Wright for helping me and data supplying and Professor Roger Strömberg for designing the compound.

I would like to give a special thanks to my supervisor in University of Minho Professor Luís Marques and to Professor Andreia Gomes, for the support they gave me.

I would like to mention the International Relations Services of University of Minho for helping me preparing my mobility programme and, consequently, making my thesis possible. Moreover, I would like to thank the funding by Erasmus+ programme of the European Union.

I am very grateful to the entire group (besides Professor Lennart and Evdokiya, also Professor Alessandra Villa, Yossa Hartono, You Xu and Arzu Uyar too) for helping me in some occasional problems and for providing suggestions and feedback to improve.

I would also like to thank many people who were always there for me, during this year. Thank them for listening, giving suggestions, moral support or just complementing my colourful plots. To keep it short, I am very thankful to my flatmates, strong princesses, friends from Portugal (in particular from Braga, Coimbra, Coja and Guimarães).

Last but definitely not least, my family. I have to thank them for providing me this opportunity and to motivate me to go further... Honestly, there are no words able to describe the amount of my gratitude.

Thank you!

Abstract

Intrinsically Disordered Proteins have regions that lack a stable structure under physiological conditions, yet they constitute an ensemble of conformations. Intrinsically Disordered Proteins are present in many biological functions, like the regulation of transcription. The altered regulation of Intrinsically Disordered Proteins is associated with many human diseases. The possibility of inhibiting Intrinsically Disordered Proteins is of great importance to chemical biology and drug discovery. Transcription Factors, like Glucocorticoid Receptor and c-Myc, are examples of proteins that contain large disordered regions. c-Myc disordered regions are associated with Burkitt's lymphoma and therefore it is considered a promising drug target.

In this work we studied, by Molecular Dynamics simulations, a peptide fragment (from residue 187 to 202) of a disordered region of Glucocorticoid Receptor. Using the same methodology, we chose a peptide fragment (from residue 42 to 63) of a disordered region of c-Myc, to study it alone and with a ligand, which represents a model drug compound against Lymphoma. The major findings from this work are:

- The data on the fragment of the Glucocorticoid Receptor showed an inverse correlation between relative activity and disorder;
- Single mutation of peptide fragments of the Glucocorticoid Receptor did not show a general trend between experimental biological relative activity and the its stability;
- The substitution of Prolines have both a destabilising and stabilising effect in peptide fragment of c-Myc;
- Hydrophobic contacts points are the most important kind of interaction between a peptide fragment of c-Myc and the ligand;
- Our data also suggest the ligand is bound to the residues located at the middle of the peptide fragment of c-Myc more specifically to the Isoleucine (I49), Tryptophan (W50) and Phenylalanine (F53);
- As future perspectives this work might constitute the basis of further evaluation of the ligand as a drug with the possibility to inhibit c-Myc activity, with great importance to for drug discovery.

Resumo

As Proteínas Intrinsecamente Desordenadas possuem regiões que, em condições fisiológicas, não têm uma estrutura estável, apesar disso formam conjuntos de conformações. As Proteínas Intrinsecamente Desordenadas participam em várias atividades biológicas, entre as quais a regulação da transcrição. A desregulação de Proteínas Intrinsecamente Desordenadas tem sido associada com várias doenças, pelo que, a possibilidade de as inibir, pode conduzir a novos fármacos e terapias. Fatores de Transcrição, como o Recetor da Glucocorticoide ou c-Myc, são exemplos de proteínas que possuem regiões desordenadas. A proteína c-Myc está associada ao linfoma de Burkitt e, como tal, é considerado um potencial alvo para novas terapias.

Neste trabalho foi estudado, recorrendo a simulações de Dinâmica Molecular, um fragmento peptídico (dos resíduos 187 a 202) de uma região desordenada do Recetor do Glucocorticoide. Utilizando a mesma metodologia, foi escolhido um fragmento peptídico (do resíduo 42 a 63) de uma região desordenada da proteína c-Myc, a fim de a estudar sozinha e com um ligando, que poderá servir como modelo de uma nova terapêutica. Dos resultados deste trabalho, destaca-se os seguintes:

- Verificou-se que a desordem do fragmento do Recetor do Glucocorticoide é inversamente proporcional à atividade biológica;
- Mutações dos fragmentos do Recetor do Glucocorticoide não mostram relação evidente entre a atividade biológica e a estabilidade;
- As substituições de Prolinas nos fragmento de proteína c-Myc demonstraram ter efeitos tanto estabilizador como destabilizador;
- Os contactos hidrofóbicos são o tipo de interação mais importante entre o fragmento da proteína c-Myc e o ligando;
- Os dados também sugerem que o ligando se liga aos resíduos localizados na porção média do fragmento da proteína c-Myc, mais especificamente à Isoleucina (I49), Triptofano (W50) e Fenilalanina (F53);
- Estes resultados perspectivam um trabalho futuro de avaliação do ligando como fármaco inibidor da actividade c-Myc, de grande importância para a descoberta de novos fármacos.

Contents

Acknowledgements i

Abstract iii

Resumo v

List of Figures ix

List of Tables xi

List of Abbreviations xiii

1 Introduction 1
- 1.1 Intrinsically Disordered Proteins . . . 1
  - 1.1.1 Amino Acids . . . 1
  - 1.1.2 Structure . . . 2
  - 1.1.3 Interactions . . . 3
  - 1.1.4 Function . . . 5
  - 1.1.5 Regulation . . . 5
  - 1.1.6 Diseases . . . 7
  - 1.1.7 Drug targets . . . 7
- 1.2 Transcription Factors . . . 7
  - 1.2.1 Glucocorticoid Receptor . . . 8
  - 1.2.2 c-Myc . . . 10
- 1.3 Molecular Dynamics Simulations . . . 12
  - 1.3.1 Force Field . . . 13
  - 1.3.2 Particles motion . . . 15
  - 1.3.3 Running the simulations . . . 16
  - 1.3.4 Accuracy vs speed . . . 17

2 Objectives 19

3 Methods 21
- 3.1 Intrinsically Disordered predictions . . . 21

3.1.1 PONDR 21
3.1.2 IUPred 21
3.1.3 ESPRITZ 22
3.2 Activity measurements 22
3.3 Molecular Dynamics Simulations 23

4 Results and discussion 25
4.1 Glucocorticoid Receptor 25
4.1.1 Mutant selection of τ_1 core 25
4.1.2 Stability in τ_1 core 27
4.2 c-Myc 32
4.2.1 MBI alone 32
4.2.2 MBI with the compound 32

5 Conclusion 37

Bibliography 39

List of Figures

1.1 Conformations of Intrinsically Disordered proteins 3
1.2 Free energy landscape 4
1.4 Functional and non-functional interactions 6
1.5 Availability and its outcomes 6
1.6 Schematic representation of Glucocorticoid Receptor 9
1.7 c-Myc functions 10
1.8 Heterodimer of c-Myc and Max and DNA 11
1.9 Schematic representation of c-Myc 12
1.10 Experimental and computational approaches 13
1.11 Lennart-Jones potential 15
1.12 Diagram of the molecular dynamic process 17

4.1 Disorder prediction of τ_1 core 26
4.2 Disorder difference as a function of the relative activity 27
4.3 Helicity of τ_1 core of Glucocorticoid Receptor 28
4.4 Frequency of t_h of τ_1 core of Glucocorticoid Receptor 28
4.5 Helicity first passage time function of the activity 30
4.6 Helical hydrogen bonds of τ_1 core of Glucocorticoid Receptor . . 30
4.7 Helical hydrogen bonds of MBI of c-Myc protein 33
4.8 Schematic figure of MBI of c-Myc and Ligand 34
4.9 Interactions between MBI region of c-Myc and the ligand 35

List of Tables

1.1 Amino acids and disorder 2

4.1 Wild type and mutants information 29

List of Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
DBD	DNA binding domain
LBD	Ligand binding domain
NTD	N-terminal domain
AF	Activation function
MB	c-Myc box
CHARMM	Chemistry at HARvard Macromolecular Mechanics
NMR	Nuclear magnetic resonance
WT	Wild type

1 Introduction

1.1 Intrinsically Disordered Proteins

In 1894, Emil Fischer recognised the lock-and-key mechanism, in which enzymes and the substrate are compared to a lock being opened by a key [1]. In this analogy, enzymes need a well-defined structure so as to bind to the substrate and consequently fulfil their function. This correlation between structure and function led to the creation of protein structure-function paradigm [1, 2].

Years later, in 1978, proteins that do not follow the paradigm were discovered [1]. Nowadays it is possible to isolate function by genetic methods as using mutants and knockouts and these methods allow identification of unfolded proteins [3]. These proteins have regions that lack stable secondary or tertiary structure but are functional [1, 2, 4–8]. Many terms were used to describe such regions like natively, naturally, inherently or intrinsically and unfolded, unstructured, disordered or denatured [1]. Currently these proteins are known as Intrinsically Disordered proteins [1, 2, 4] and their research has grown extensively [3].

1.1.1 Amino Acids

Before the discovery of Intrinsically Disordered proteins, in 1973, Christian Anfinsen proved the amino acid sequence of a protein encodes the information for its folding [5]. In the case of globular proteins, which are rich in hydrophobic amino acids, fold into a hydrophobic core [7]. On the opposite case, for Intrinsically Disordered proteins, their sequence explains their inability to fold [5, 6, 8, 9] and, in its regions, amino acids are often charged or polar, whereas hydrophobic are scarcer [6, 7, 9–13]. See table 1.1. Therefore they lack a hydrophobic core and the charge distribution destabilises the structure affecting its extension [6, 11, 12]. Furthermore, it was suggested the amino acid composition of Intrinsically Disordered regions might be more important than the specific sequence itself [2]. Amino acids inside a protein or a peptide fragment are mentioned as residues and this terminology will be used here.

Table 1.1: This table links amino acids and disorder. The first part describes the amino acid composition of Intrinsically Disordered regions. The second part ranks the amino acids according to their disorder properties.

Intrinsically Disordered Regions [6]	
Rich in	Arg, Gln, Glu, Lys, Pro, and Ser
Deficient in	Cys, Ile, Leu, Phe, Trp, Tyr, and Val
Similar levels with folded proteins	Met and Thr
Enriched in some and depleted in others	Ala, Asn, Asp, Gly and His

From order to disorder promoting [7]
Trp, Phe, Tyr, Ile, Met, Leu, Val, Asn, Cys, Thr, Ala, Gly, Arg, Asp, His, Gln, Lys, Ser, Glu and, at last, Pro

1.1.2 Structure

Intrinsically Disordered regions can be just a few residues long, extensive loops or, usually, protein ends [6, 7]. Since Intrinsically Disordered regions lack an hydrophobic core, they do not fold spontaneously into the compact states as the ones found in the Protein Data Bank [13]. That is, Intrinsically Disordered Proteins have regions that lack stable secondary or tertiary structure under physiological conditions [1–12, 14]. Instead, Intrinsically Disordered proteins are dynamical ensembles of flexible interconverting conformational sub-states with low energy barriers [1, 3, 5, 8, 10, 12]. These sub-states cannot be described by the position of their atoms and backbone angles, as they do not define a unique structure[1, 5, 8].

Even so, Intrinsically Disordered proteins are never completely random coils, they are also likely to generate regions of secondary structure [13]. The amount of this secondary structure varies and, depending on it, Intrinsically Disordered proteins turn out more compact or extended [5]. If it is extended, is described as being in random coil [5, 7, 13], which means the coil has random backbone angles [13]. See figure 1.1a. In case an Intrinsically Disordered protein exhibits a more compact conformation, it is molten globule [5, 7, 13]. See figure 1.1b. If Intrinsically Disordered proteins actually have globular domains divided by Intrinsically Disordered linkers, they are modular, as beads on a string [13]. See figure 1.1c. Yet these flexible linkers, when bound, achieve a structure [13]. Hence Intrinsically Disordered regions structure not only depends on its amino acid sequence, but also on its environment [8].

(a) Random coil (b) Molten globule (c) Linked folded domains (like beads on a string)

Figure 1.1: Representation of different conformations of Intrinsically Disordered proteins. Adapted from [13].

1.1.3 Interactions

In general, upon interacting with a ligand, Intrinsically Disordered proteins normally fold into a well-defined structure, meaning they go through coupled binding and folding [2–6, 13]. While the unbound protein does not have a stable structure, the complex has [6]. See figure 1.2. The folding process corresponds to a balance between the decrease of the entropy with disorder-to-order transition and the decrease of enthalpy [3, 5, 6, 13, 14]. The entropic cost enables Intrinsically Disordered proteins to interact with high specificity and low affinity [2, 4, 11, 13], so their interactions are precise but weak and brief [10, 11]. In other words, this allows them to associate and dissociate quickly and easily [2].

Besides, due to the structure extension, Intrinsically Disordered proteins have a wider interaction surface area compared to globular proteins of the same length [4]. Intrinsically Disordered proteins bind to several partners namely proteins, such as themselves, nucleic acids and others [2, 8]. Additionally, their interactions may include the binding of many different regions to one partner, see figure 1.3a, and one region to many different partners, see figure 1.3b [8]. This second one is possible because Intrinsically Disordered proteins have binding sites in quick succession and fold differently depending on the ligand [2, 3, 8]. Given these multiple binding possibilities, it is possible to infer that ligands may compete for the binding [8].

Along with coupled binding and folding, Intrinsically Disordered proteins also experience allosteric effects. Allostery is a change of structure induced by the binding to a first ligand, which alters the binding affinity to a second ligand [15]. The interference of allosteric effects cannot be disregarded in Intrinsically Disordered proteins. For example, a certain domain may have an allosteric response to the binding to a ligand of another domain. As a consequence, this domain can have its binding affinity increased/decreased. Being so, a specific ligand behaves as an agonist/antagonist depending on the previous state in the ensemble [9].

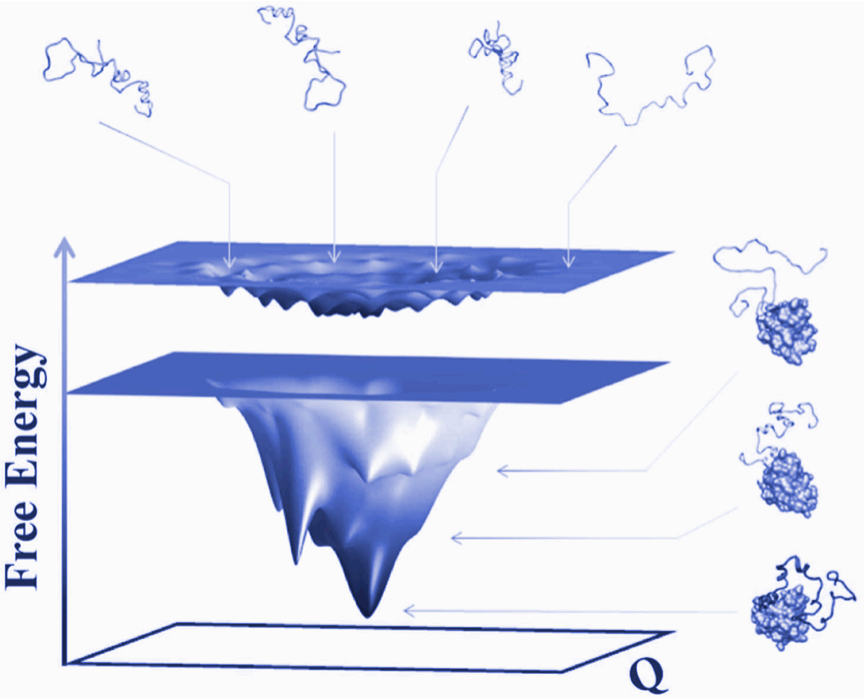
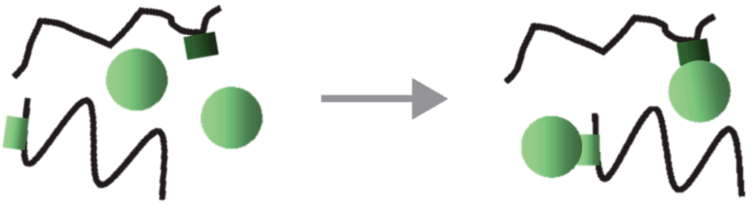


Figure 1.2: The portrayal of the free energy landscape: on the top, the free energy of an unbound Intrinsically Disordered protein; on the bottom, the free energy of Intrinsically Disordered protein binding. The unbound protein is illustrated by a flat surface, in contrast with the bound protein, which shows a distinct minimum. Adapted from [14].



(a) The sketch shows two Intrinsically Disordered protein with different binding regions, small dark and light green boxes, and a ligand, light green circles. The picture illustrates the binding of many Intrinsically Disordered regions to one ligand.

(b) The sketch shows an Intrinsically Disordered protein with its binding regions, small dark green boxes, and two different ligands, dark and light green circles. The picture illustrates the binding of one Intrinsically Disordered region to many ligands.

Figure 1.3: Representations of Intrinsically Disordered proteins binding to their ligands. Adapted from [4].

1.1.4 Function

Intrinsically Disordered proteins have a range of functions, which complement the functions of ordered proteins [7]. Regarding Intrinsically Disordered proteins, the broader significance of their "structures" benefits their function [1, 3]. Such advantages include binding to numerous partners and being involved in different interaction pathways as hubs [3, 4]. Also, their interactions have fundamental features for signalling functions. Signalling processes are initiated with high associating specificity; when a process is finished, the low affinity allows a quick dissociation [4, 13]. Beyond signalling, Intrinsically Disordered proteins are known for cell cycle control, regulation of transcription and translation, recognition of DNA, RNA and other proteins, supporting in the folding proteins and RNA, determining cell's response to an external stimulus [1, 2, 6–8, 10, 11]. Conversely, Intrinsically Disordered proteins are usually not involved in control of metabolism, biosynthesis, transport and other functions that entail a specific structure [1, 11].

1.1.5 Regulation

Tight regulation of Intrinsically Disordered proteins enables fidelity in their functions. This control is provided at various levels [4, 5] and it aims to preserve equilibrium, as the proper concentration, post-translational modifications, position and lifespan in the cell [4].

Regulation of the availability

As an Intrinsically Disordered region can bind to many different partners and different regions can bind to the same partner (described before in 1.1.3), it is suggested they are capable of binding to each other's partners and also that these interactions are more likely to occur in a higher concentration of Intrinsically Disordered proteins [4]. See figure 1.4. Such a promiscuous behaviour may cause non-functional interactions, destabilising signalling pathways, thus leading to undesirable outcomes. Therefore, Intrinsically Disordered proteins regulation controls their availability, by being present in proper levels and no longer than required, which generally results in lower amounts and for shorter intervals than ordered proteins [4]. See figure 1.5. In spite of that, not every Intrinsically Disordered protein in a situation of increased expression is harmful. In certain circumstances as being under stress, in a precise period of the cell cycle or in reaction to certain stimuli, could require a higher availability [4].

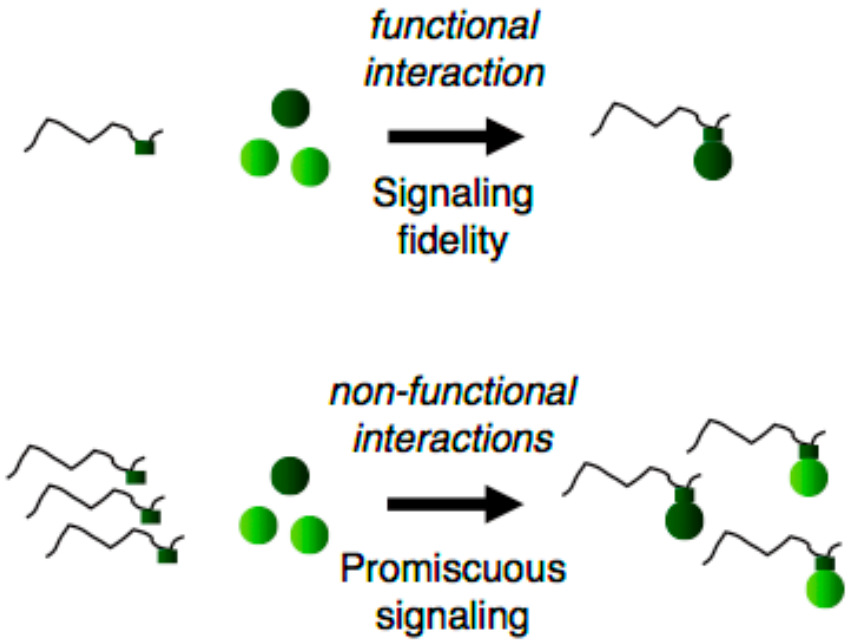


Figure 1.4: The sketch shows an Intrinsically Disordered protein with its binding region, small dark green box, and two different ligands, dark and light green circles. The top illustrates the proper availability, a functional interaction is achieved. The bottom illustrates an increased availability, not only functional interaction but also non-functional interactions are achieved. Adapted from [4].

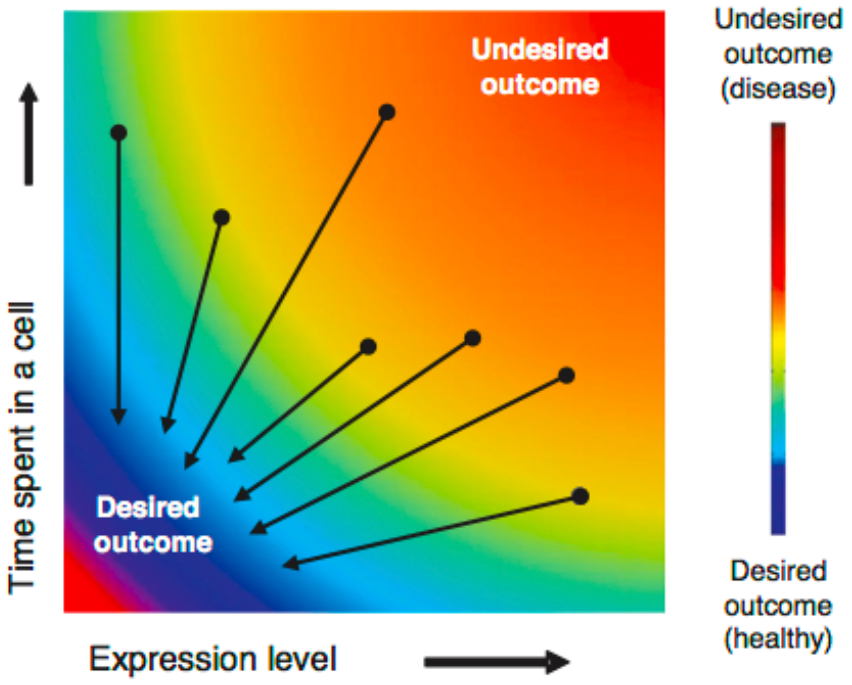


Figure 1.5: Hypothetical "Availability-outcome" landscape of an Intrinsically Disordered protein. Proper availability generates a desired outcome (blue region); altered availability generates an undesired outcome (red region). Adapted from [4].

1.1.6 Diseases

Intrinsically Disordered proteins have been involved in pathological mechanisms of human diseases, in more than one way. In one hand, the concept of controlled-chaos arose from the regulation of Intrinsically Disordered proteins, being the uncontrolled-chaos recurrently connected with diseases [5]. On the other hand, the amino acid composition and its tendency to develop beta-structures, such as β-sheets and β-turns, has been described as another mechanism. These cause higher aggregation potentials with propensity to create structures of the amyloid kind [10]. Thus Intrinsically Disordered proteins are found related with diseases mainly by loss of regulation or aggregation effect: different cancers, Alzheimer's, Parkinson's, cardiovascular diseases, diabetes, neural diseases, prion diseases, accelerated fibrillation, protein deposition diseases, among others [1, 5, 10, 11]. Because of the connection to these illnesses, Intrinsically Disordered proteins are becoming potential drug targets, where the target is the Intrinsically Disordered protein itself or the protein-protein interactions [10].

1.1.7 Drug targets

Recent medicines treatments only have made use of less than 10% of reasonable potential targets, which are exclusively around 500. Besides, roughly 70% of these drug targets are restricted to membrane receptors and enzymes [16].

Moreover, it was mentioned healthcare and pharmaceutical industry suffered serious drop of income [17]. This was due to further demanding requirements for regulating procedures, growing cost-constrained healthcare systems and patent expirations. To overcome this situation, they seek new additional ground breaking, reasonable priced and effective drugs [17].

For the new drug molecules, targeting protein-protein interactions is a likely approach. As mentioned, Intrinsically Disordered proteins bind to proteins, so they establish such described interaction. This kind of interaction produces a response upon external stimuli. Blocking it with a small molecule is an attractive therapeutic intervention. The possibility to inhibit Intrinsically Disordered proteins is of great importance to chemical biology and drug discovery [16].

1.2 Transcription Factors

Transcription Factors are an example of proteins that contain large disordered regions [18]. Their function is to control gene expression by recognising and binding to specific DNA sequences [19, 20]. Transcription Factors identify around 6 to 12

base pairs and bind to regulatory regions [20, 21]. These proteins regulate through activating or repressing transcription [19] and, when inducing transcription, Transcription Factors operate together with RNA polymerase, co-factors and others [22]. Transcription Factors are influenced by several factors such as the nucleus activity, accessible DNA binding sites and the activation previous state [9, 19].

1.2.1 Glucocorticoid Receptor

Glucocorticoid receptor is a member of a family of nuclear receptors [23–25]. Glucocorticoid Receptor is a modular protein, so each domain is capable of preforming distinct functions [23, 26]. In general, nuclear receptors can be divided in ligand binding domain (LBD), DNA binding domain (DBD), hinge region and N-terminal (NTD) [9, 27]. See figure 1.6 and table 1.2.

As the name implies, Glucocorticoid Receptor is a receptor of glucocorticoids [24, 26], which are steroid hormones (cortisol, for example), and they are produced in the adrenal cortex [28]. Glucocorticoids are involved in immunosuppressive and anti-inflammatory functions, gluconeogenesis, lipolysis and decomposition of proteins [28]. Inactive Glucocorticoid Receptor are located in the cytoplasm, in a complex with heat shock proteins [24, 26]. Once glucocorticoids bind to these Glucocorticoid Receptors, they dissociate from the complex with heat shock proteins and turn active. Following this, the receptor and its ligand move into the nucleus and bind as a homodimer to specific DNA sequences [23–25]. These DNA sequences, called glucocorticoid response elements, are located in the surrounding of the target genes and regulate their activity [24, 25, 27].

An activation or transactivation domain is a small region of the protein with the capacity to activate transcription [29]. In nuclear receptors there are at least two activation domains. One, the activation function 1 (AF1), is located in the NTD and it is ligand-independent. The other, activation function 2 (AF2), is positioned in the LBD and it is ligand-dependent [9, 29]. In humans, AF1 is more active and situated between residues 77 and 262. In opposition, AF2 is a minor domain and located between residues 526 and 556 [23–25]. There is a third conserved region in the C terminus that may participate in the transactivation activity [23, 24]. Regarding AF1, also known as τ_1, it encloses a region with near 60% or 70% of the activity of the entire domain. This region is called the τ_1 core and is a sequence of 58 amino acid from 185 residues of τ_1 [23–25]. See figure 1.6. Even though its function, τ_1 and τ_1 core were proven to be unstructured in aqueous solution [24].

Table 1.2

Ligand Binding Domain	Domain where the typical small ligands bind; domain conformation and protein activity varies with the ligand bound;[9]
DNA Binding Domain	Domain sensitive to DNA sequences of its response elements; response elements are a specific DNA sequence where the protein binds;[9]
N-terminal Domain	Domain with high Intrinsically Disordered content; domain with allosteric effects on LBD and DBD;[9]
Hinge region	Domain that links DBD and LBD;[27]

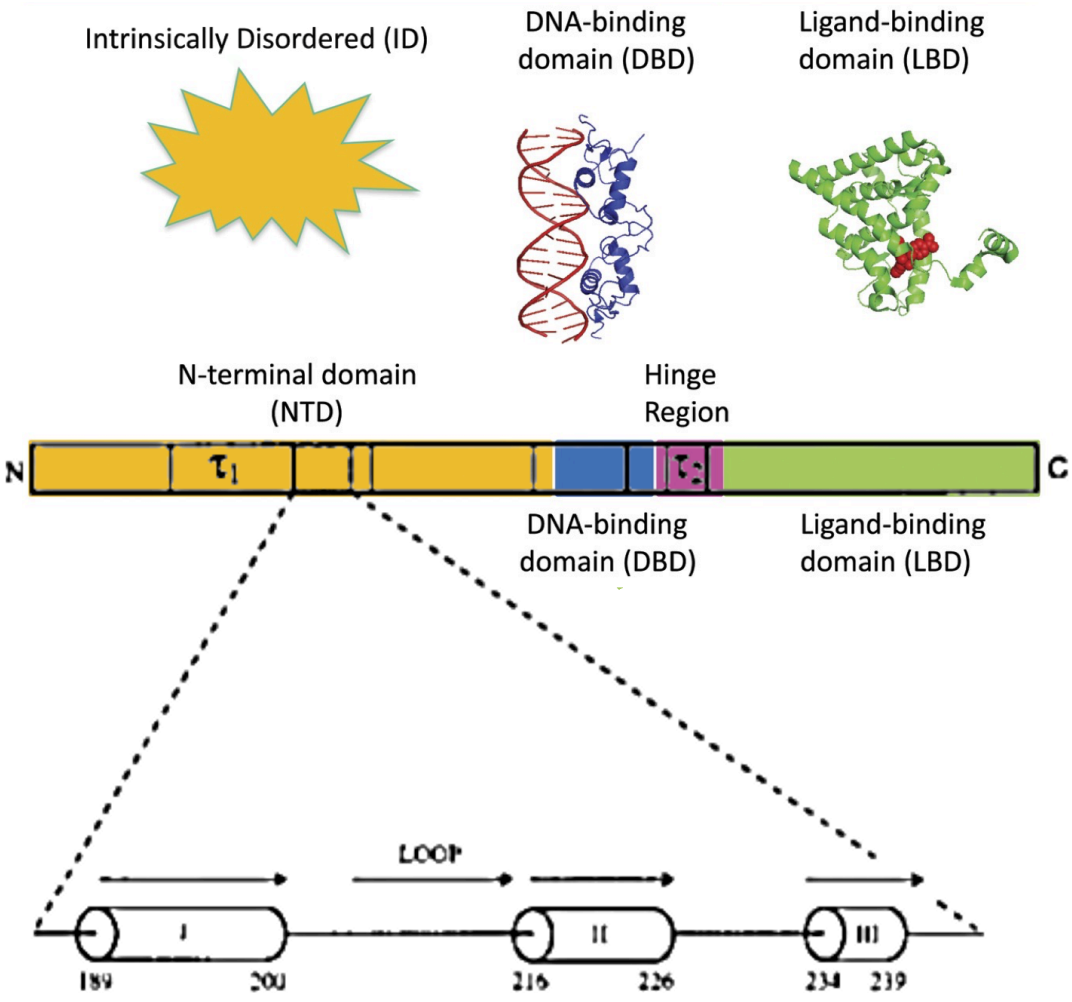


Figure 1.6: Schematic representation of the human Glucocorticoid Receptor and of its the τ_1 core transactivation domain. Model is constituted by globular ligand binding domain (green), globular DNA binding domain (blue), hinge region (pink) and disordered N-terminal domain (yellow). The locations of the α-helixes are shown. The sequence of τ_1 core (HRI) is displayed. Adapted from [9, 23].

1.2.2 c-Myc

c-Myc, in short Myc, is a transcriptional factor that regulates several biological functions. These include cell growth, cycle and survival, metabolism and energy production, proliferation, differentiation and apoptosis [30–34]. c-Myc is an example of a hub by assimilating various pathway signals and controlling many functions [32]. See figure 1.7. c-Myc levels are highly regulated at MYC gene expression, protein stability and others [32]. Nonetheless these turn out deregulated in many cancers [30, 32, 33]. For example, in Burkitt's lymphoma, MYC gene is frequently found mutated [32]. c-Myc abnormal expression may lead to a cell transformation, proliferation and resistance to apoptosis in hypoxia, external stress and other hostile factors of microenvironments [33]. With such noticeable association to cancer, c-Myc was considered a promising drug target [32].

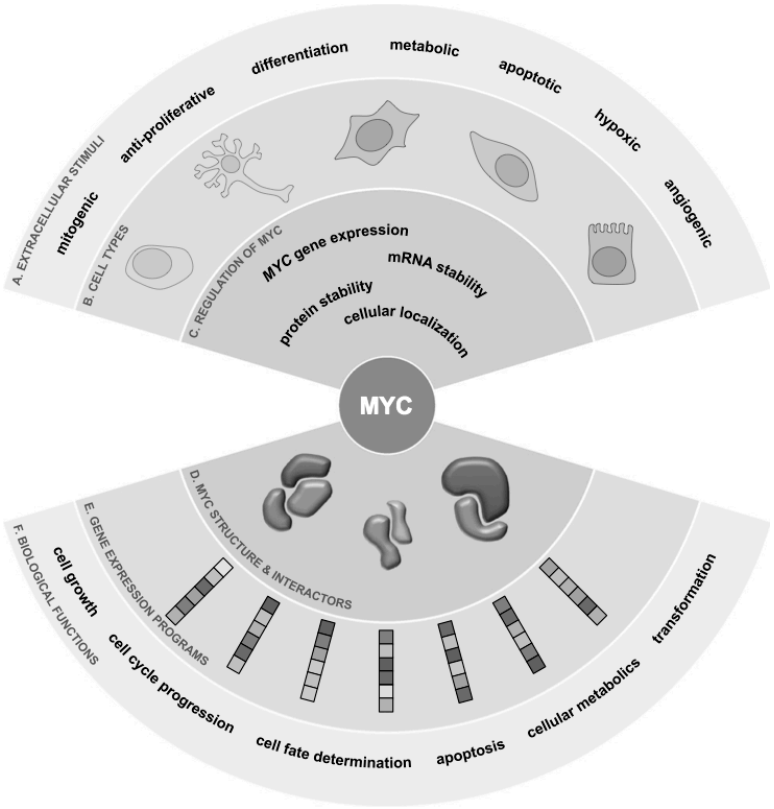


Figure 1.7: Representation of the biological functions associated with c-Myc in the cell, acting as a central hub. Adapted from [32].

C-Terminal

c-Myc has a basic helix-loop-helix (bHLH) DNA binding and a leucine zipper (LZ) dimerisation motifs. c-Myc heterodimerises with its partner protein Max in order to bind DNA with CACGTG sequence [30–32]. See figure 1.8. Max is essential for c-Myc's transcriptional activities, therefore essential for c-Myc's function [32].

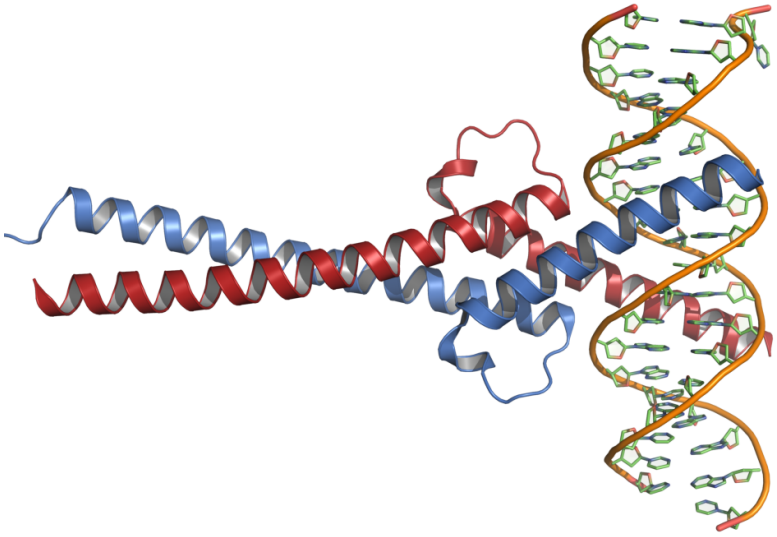


Figure 1.8: Heterodimer of c-Myc (red) and Max (blue) proteins binding to DNA. Adapted from [35].

N-Terminus

In the N-terminus of c-Myc, there are regions with high sequence conservation between species and these regions are called Myc Boxes (MB) [32]. The first, MBI is positioned between the residues 47 and 63 [31]. MBI is able to regulate c-Myc by its two phosphorylation sites: T58 and S62 [31, 32]. See figure 1.9. The phosphorylation of S62 activates c-Myc and the sequential phosphorylations of S62 and T58 are initial steps of c-Myc degradation pathway [32, 34]. Mutations around T58 and S62 are connected to cancers as Burkitt's lymphoma [30–32, 34]. These interfere with T58 phosphorylation, which gather and hold c-Myc activated in S62-phosphorylated state [31, 34]. Hence, MBI is fundamental to set c-Myc activity time frame (c-Myc half-life is 20-30min) [31].

Even though N-terminal is considered disordered, there are two regions of transient secondary structure. One is located from the amino acid 22 to 33 and the other region overlaps MBI. This last region, has two segments: between residues 48 and 55, with helical properties, and between residues 56 and 65, with fluctuating extended character possibly stabilised by a proline enriched sequence.[31]

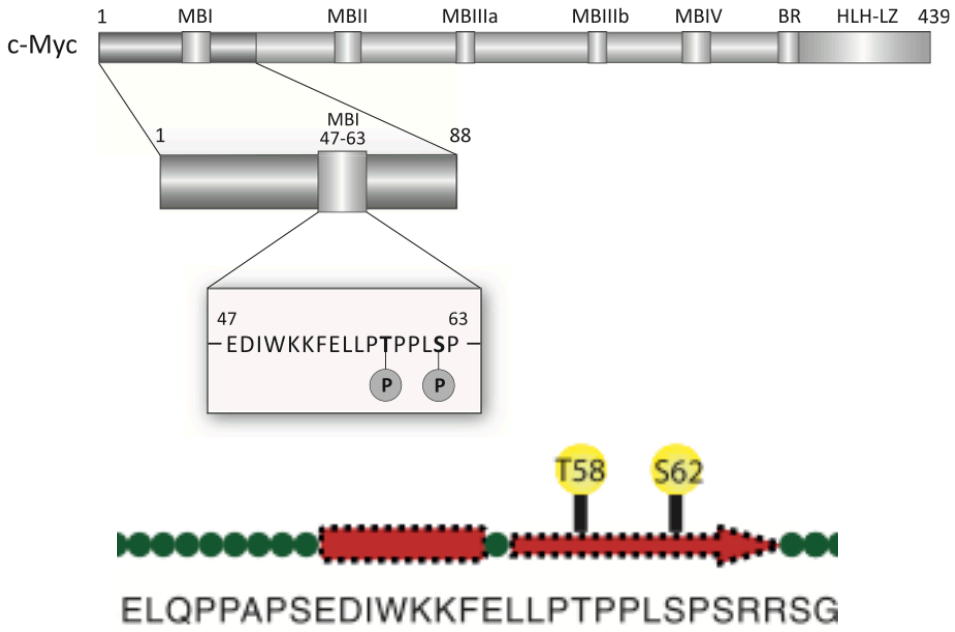


Figure 1.9: Schematic representation of c-Myc protein, location of MBI domain and its phosphorylation sites. It is also identified the region with helical properties (red box) and between the region with fluctuating extended character (red arrow). Adapted from [31].

1.3 Molecular Dynamics Simulations

"Certainly no subject or field is making more progress on so many fronts at the present moment than biology, and if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms."[36]

Molecular dynamics is a valuable tool that provides information of a system by simulating the motion of particles over time [37]. In here, particles usually mean atoms. Specifically, it informs about conformations, system in equilibrium and thermodynamic parameters [38]. When applied to a biological system, like proteins, improves the understanding of the biological phenomena [37].

Compared to experiments, this computational approach has advantages, for instance being easier to measure certain properties. Nonetheless experimental approaches allow assessing the accuracy of the simulations results and improving methodology [38]. Actually, the computational and experimental approaches should be regarded as complementary techniques. See figure 1.10.

Programs like CHARMM, Amber and GROMACS are examples of molecular dynamics simulations programs [39, 40].

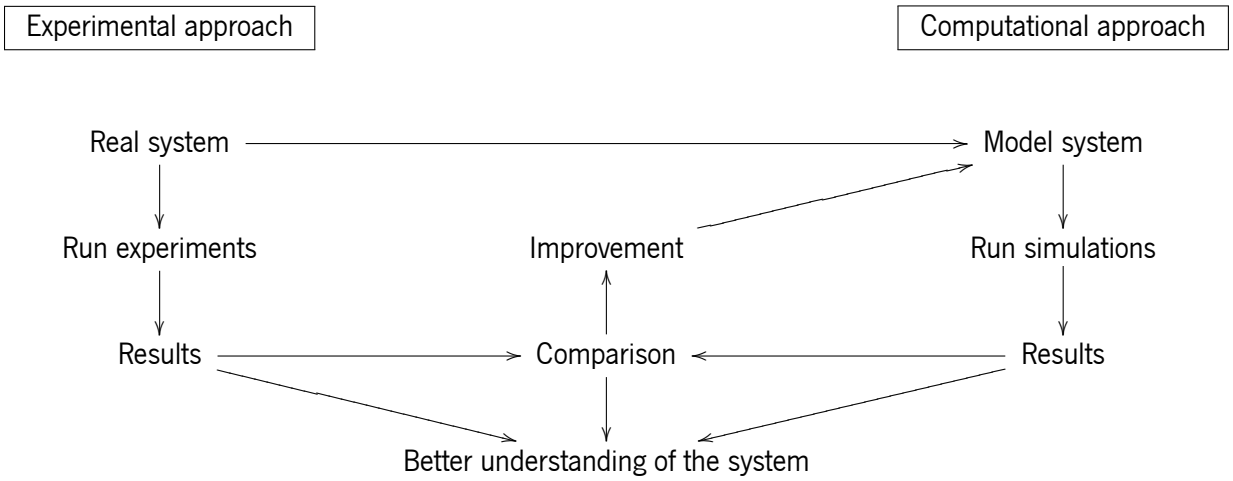


Figure 1.10: Interactions between experimental and computational approaches. Adapted from [41]

1.3.1 Force Field

In order to calculate the particles motion, a potential energy function must be defined [37, 39]. It is through this potential that forces can be calculated and subsequently obtained the equations of motion. For simple systems of few and small molecules, the potential energy function could be achieved by quantum mechanics calculations. However, most systems, like proteins or other macromolecules, are not so simple. They require an empiric potential function, which is based on classical mechanics of fixed point charges [37, 42]. The potential energy function is also called force field. Some examples of force fields are CHARMM22 force field for proteins, CHARMM27 force field for nucleic acids, AMBER nucleic acid, OPLS force field [42].

Typically the potential energy function U has two contributions: interaction of bonded and non-bonded atoms [37, 40, 42]. See (1.1). The bonded interaction contribution ($V_l+V_\theta+V_\phi$) to the force field is characterised by the length, angles and dihedral angles of bonds between atoms. The non-bonded contribution is constituted by the electrostatic potential $V_{Coulomb}$ and the Lennart-Jones potential V_{LJ}.

$$U = V_l + V_\theta + V_\varphi + V_{Coulomb} + V_{LJ} \tag{1.1}$$

Bonded contibutions

The components V_l and V_θ are represented by harmonic potentials, where the length l and angles θ, respectively, correspond to a displacement x [37, 39, 40, 42]. They correspond to the stretching and twisting of bonds. In the harmonic potential

function as V_x, x_0 represents the equilibrium value. See (1.2). The component V_φ is a periodic potential, so the potential is expressed as sinusoidal function of dihedral angles φ [37, 39, 40, 42]. See (1.3). In here, δ is the phase shift and n is the periodicity of φ.

$$V_x = \sum_x K_x (x - x_0)^2 \tag{1.2}$$

$$V_\varphi = \sum_\varphi K_\varphi \left[1 + \cos\left(n\varphi - \delta\right)\right] \tag{1.3}$$

Non bonded contibutions

The non-bonded component is the sum of electrostatic potential v_{ij} and the Lennart-Jones potential u_{ij} of each non-bonded pairs of atoms i and j [37, 40, 42]. See (1.4).

$$V_{Coulomb} + V_{LJ} = \sum_{ij} \left(v_{ij} + u_{ij}\right) \tag{1.4}$$

The electrostatic potential v_{ij} represents the interaction between the charges q_i and q_j. This contribution can be attractive or repulsive, according to the charges q_i and q_j signs. See (1.5) [37, 40, 42].

$$v_{ij} = \frac{q_i q_j}{4\pi\varepsilon_r r_{ij}} \tag{1.5}$$

The Lennard-Jones potential u_{ij} is a function with both repulsive and attractive terms and also forms a minimum. See (1.6). The repulsive term is due to of core-core repulsion and the attractive is due to van der Waals interaction. Here, σ defines $u(r = \sigma) = 0$ and ϵ is $u(r = r_{min}) = \epsilon$, where r_{min} is the distance at the minimum potential.[37, 40, 42, 43] See figure 1.11.

$$u_{ij} = 4\epsilon \left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] \tag{1.6}$$

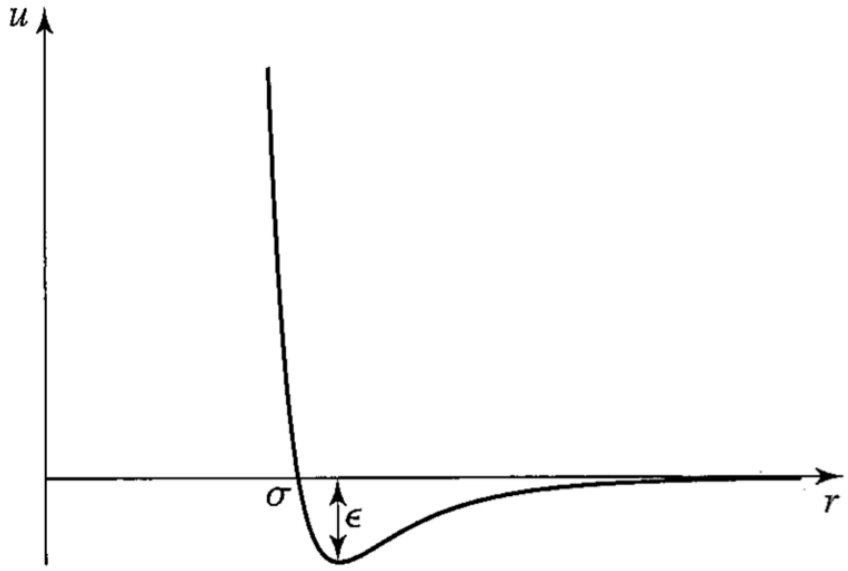


Figure 1.11: Graph showing Lennard-Jones potential. Adapted from [43].

Parameters

The force field, besides being a function, also consists of parameters. For instance, K_x, x_0 K_φ, n, δ, ε_r, ϵ, σ are parameters. The parameters K_x and K_φ are force constants. ε_r is the relative dielectric constant. These are obtained by quantum mechanics calculations or experimental methods [37, 40, 42].

1.3.2 Particles motion

Having a potential energy function, it is possible to determine the forces established between the atoms and subsequently achieve the particles motion [39, 40]. By the definitions of work W in (1.7) and potential energy U in (1.8), the force $\vec{F}$ is the negative of the gradient of potential energy function. See (1.9).

$$W = \int_c \vec{F} \cdot d\vec{r} \tag{1.7}$$

$$W = -\Delta U \tag{1.8}$$

$$-\vec{\nabla} U = \vec{F} \tag{1.9}$$

According to the second Newton's law, the velocity $\vec{v}$ of a particle can be calculated by the temporal integration of its force $\vec{F}$ and the position $\vec{r}$ by a double

temporal integration. The constant m is the mass of the particle. See (1.10).

$$\vec{F} = m\frac{d^2\vec{r}}{dt^2} \tag{1.10}$$

In molecular dynamics simulations, it is used algorithms called dynamics integrators in order to integrate force [42]. An example of a dynamics integrator is Verlet algorithm, which is based on Taylor series expansions [39]. See (1.11). Here $r(t)$ is the position in a single dimension at the instant t and Δt is the time interval of the between the previous and following position.

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + \frac{F}{m}\Delta t^2 + O(\Delta t^4) \tag{1.11}$$

1.3.3 Running the simulations

The figure 1.12 is a diagram of the molecular dynamic process to run a simulation. Molecular Dynamics simulations require a initial set of the structure. This can be given by experimental methods or by building a model [40]. Generally, the structure could be imported form a PDB file. Then the force field and its characteristic must be chosen [42].

Afterwards, it is necessary to optimise the system structure. The optimisation uses an iterative minimisation algorithm. One of the simplest method is steepest descent. Generally, more than one algorithm can be employed, for instance the Adopted-Basis set Newton-Raphson method is performed following the steepest descent method.[42]

With the structure refined, the system is ready for a heating process. At first, atoms are assigned random velocities from a maxwellian distribution associated with a low temperature. Then, each time the temperature is increased, a new set of random velocities from the distribution are assigned. The relation between the mean velocity $\langle v_i \rangle$ of the atom i and temperature T is given by equation (1.12) [37]. Here, N is the total number of the atoms, m_i is the mass of the atom i and k_B is the Boltzmann constant.

$$\frac{1}{2}\sum_{i=1}^{N} m_i \langle {v_i}^2 \rangle = \frac{3}{2} N k_B T \tag{1.12}$$

When the temperature stabilises, atoms have random velocities from the corresponding maxwellian distribution and the whole system has the same average temperature. Then all is set to run the simulations as a function of time: continue integrating the equations of motion, generating coordinates and velocities time-dependent [37].

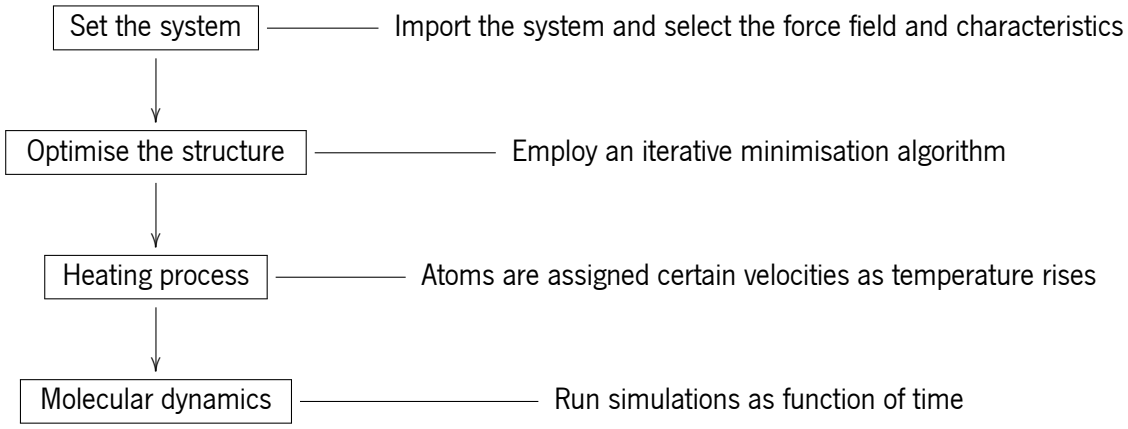


Figure 1.12: Diagram of the molecular dynamic process. On the left, there are four essential steps in order. On the right, there are descriptions of each step. From [37]

1.3.4 Accuracy vs speed

Molecular Dynamics simulations poise between accuracy and speed.[42] For example, a more accurate force field can lower the speed. So the force field must be adapted to the system for a better performance, which means more or less contributions are taken into account. An example of additional contribution is the backbone dihedral angle correction term. It corrects minor systematic errors in the protein backbone description. This term is based on quantum mechanical calculations and structure-based potentials of mean force [42].

Additionally, the non-bonded contributions can be only calculated for distances within a cutoff limit. This is to reduce numbers of non-bonded pairwise interactions taken into account and speeds the calculations [42]. When cutoff is used, it is advised to also use the force-shift methods. They smooth the forces near the cutoff distance, in order to go to zero in a continuous manner [42].

Since the environment has a great impact on the protein behaviour, proteins are frequently solvated, thus being important to simulated a solvent.[37, 42] Once again, there are options of how to represent the solvent: explicit or implicit. In explicit solvent, water molecules are simulated and the system is composed of many more atoms; in implicit solvent, the relative dielectric constant ε_r becomes distance-dependent [38, 42].

Other concept to consider in molecular dynamics is the system boundary conditions. One type of boundary conditions is the periodic boundary conditions. The point is to simulate an infinite system by repeating the central cell as if there were

no actual boundaries, in order to minimise the effect of the boundary on the system [42].

There is the possibility to apply SHAKE methods to impose holonomic constraints too. These constraints are able to fix the hydrogen bonds length, when they can be disregarded. SHAKE usage allows a longer timestep in the integration, like 2 fs [42].

2 Objectives

In this work, it is proposed to:

1. Improve the understanding of stability of Intrinsically Disordered regions of proteins like Glucocorticoid Receptor and c-Myc;
2. Evaluate a novel compound, which could be a potential drug against Lymphoma.

For that we used CHARMM to look at the stability of disordered regions of Glucocorticoid Receptor (from residue 187 to 202) and c-Myc (from residue 42 to 63). In c-Myc we further studied their interactions with a ligand which is a model drug compound, representative of a new therapeutic intervention.

3 Methods

3.1 Intrinsically Disordered predictions

To look at the level of disorder of the Glucocorticoid Receptor, we used different predictors. These predictors analyse the amino acid composition, the physicochemical properties of the sequence and the possibility of secondary structure formation. Predictors derive from machine learning and statistical approaches. We used different Intrinsically Disordered predictors, such as PONDR, IUPred and SPRITZ. Calculations done by Professor Anthony Wright.

3.1.1 PONDR

PONDR predicts upon single sequences, using neural networks that use sequence attributes taken over windows of 9 to 21 amino acids. These attributes, such as the fractional composition of particular amino acids, hydropathy, or sequence complexity, are averaged over these windows and the values are used to train the neural network during predictor construction [44]. Those predictors are all well described in literature [45].

PONDR_fit, PONDR_vL3, PONDR_vls2, PONDR_vltx are different predictors used in this work, which differentiate from each other on the algorithms used and differences conditions like coordination number, peptide size [44].

3.1.2 IUPred

IUPred is also a web-server estimating the capacity of polypeptides to form stabilising contacts. The underlying assumption is that globular proteins make a large number of inter-residue interactions, providing the stabilising energy to overcome the entropy loss during folding. Taking a set of globular proteins with known structure, it has developed a simple formalism that allows the estimation of the pairwise interaction energies of these proteins. It uses a quadratic expression in the amino acid composition, which takes into account that the contribution of an amino acid to order/disorder depends not only its own chemical type, but also on its sequential environment, including its potential interaction partners [46].

IUPred_long

IUPred_long (long disorder) used in this work encompasses at least 30 consecutive residues of predicted disorder. For this application the sequential neighbourhood of 100 residues is considered [47].

IUPred_short

UPred_short (short disorder) is suited for predicting short, probably context-dependent, disordered regions, such as missing residues in the x-ray structure of an otherwise globular protein. For this application the sequential neighbourhood of 25 residues is considered. As chain termini of globular proteins are often disordered in x-ray structures, this is taken into account by an end-adjustment parameter which favours disorder prediction at the ends [47].

3.1.3 ESPRITZ

ESPRITZ is a web server using two specialised binary classifiers both implemented with probabilistic soft-margin support vector machines. The long disorder classifier is trained on a subset of non redundant sequences known to contain only long disordered protein fragments (more than 30 amino acids). The short disorder classifier is trained instead on a subset of non redundant sequences with only short disordered fragments [48].

Predictors used in this work use data from NMR (Espritz_nmr) and from x-ray spectroscopy (Esprits_xray) data.

3.2 Activity measurements

Relative activities were previously obtained by phenotypic screening (data in table 4.1, supplied by Professor Anthony Wright). The biological activity of the Glucocorticoid Receptor wild type and their τ_1 core mutants were measured by β-galactosidase activity expressed the yeast $Saccharomyces\ cerevisiae$ [23]. Relative activity A_{rel} corresponds to fraction of the mutant activity A_{mut} and the wild type activity A_{WT}. See (3.1).

$$A_{rel} = \frac{A_{mut}}{A_{WT}} \times 100 \tag{3.1}$$

3.3 Molecular Dynamics Simulations

The Molecular Dynamics simulations were performed with the program CHARMM using the force field CHARMM36, except for the compound, using the CGenFF. The model used is an atomistic representation and the original structures were built placing the residues in α-helix conformation. We simulated the regions from the residue 187 to 202 (τ_1 core) of Glucocorticoid Receptor and from residue 42 to 63 (MBI) of c-Myc. The simulations were in a cubic box with periodic boundary conditions of 64Å side for Glucocorticoid Receptor and 72Å side for c-Myc. The box side corresponds roughly to the protein size with an increase of 40Å. This increment is also due to Intrinsically Disordered proteins having a more extended structure. The system was then minimised with 1000 steps of steepest descent algorithm. We also used harmonic restraints in the peptide backbone, SHAKE algorithm on bonds containing hydrogens and 11Å cutoff. Afterwards, it was heated from 110K to 330K, 360K and to 400K. The dynamics integrator used is the standard integrator LEAP, which is based in Verlet leap-frog algorithm, with a 0.002ps timestep.The simulations were run in constant pressure and temperature (isothermal-isobaric system). For each system, 10 replicas were simulated for a maximum of 100ns for the τ_1 core of Glucocorticoid Receptor and 50ns for MBI of c-Myc protein and these were carried out in a Graphics Processing Unit of Karoliska Institutet cluster.

The occupancy (range between 0 and 1) is the average number of hydrogen bonds or hydrophobic contacts formed during the trajectory. This measurement was used in the simulations of MBI region of c-Myc protein and the ligand to infer about their interaction.

4 Results and discussion

4.1 Glucocorticoid Receptor

4.1.1 Mutant selection of τ_1 core

Previous work done by Professor Anthony Wright is briefly reported here.

Different predictors (described in the previous chapter in 3.1) were used to access the levels of structural disorder. In figure 4.1, it is shown the extend of disorder with the τ_1 region of the Glucocorticoid Receptor. This region is known to be unstructured in aqueous solutions and to have three regions with helical propensity structures by NMR [23]. The helical propensity region H1 is the described τ_1 core, which contains nearly 60% to 70% of the biological activity of the entire domain [23]. All these facts are the major reason of choice to perform further studies of the first helical structure (τ_1 core) of Glucocorticoid Receptor.

Experimental data of the relative biological activity of mutations within the τ_1 core region (experimental data from literature [23]) and the disordering level are shown in figure 4.2. It clearly shows an inverse correlation between disorder and the relative activity of mutants. The relative activity A_{rel} is calculated relatively to wild type, described in the previous chapter in 3.2.

In this work, fourteen mutated peptide fragments of the τ_1 core region (residues 187 to 202) were selected from the data of figure 4.2 by presenting a smaller deviation from the inverse correlation between relative activity and disorder difference of the τ_1 core of Glucocorticoid Receptor.

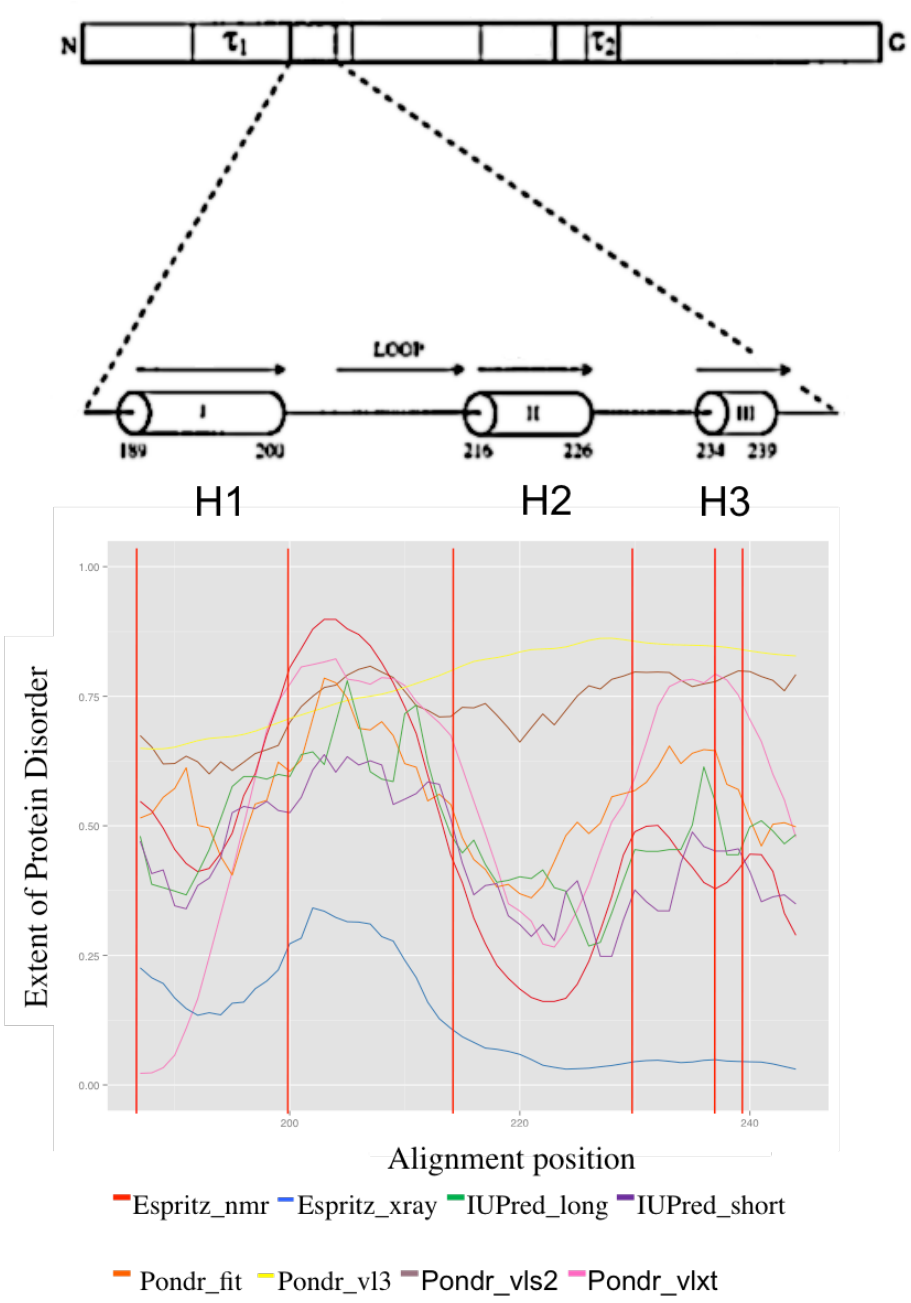


Figure 4.1: On the top, schematic representation of Glucocorticoid Receptor and its three helical propensity regions (H1, H2 and H3) by different predictors. On the bottom, extent of protein disordered of the mentioned regions by different predictors. The predictors are described in the previous chapter in 3.1. Data kindly supplied by Professor Anthony Wright.

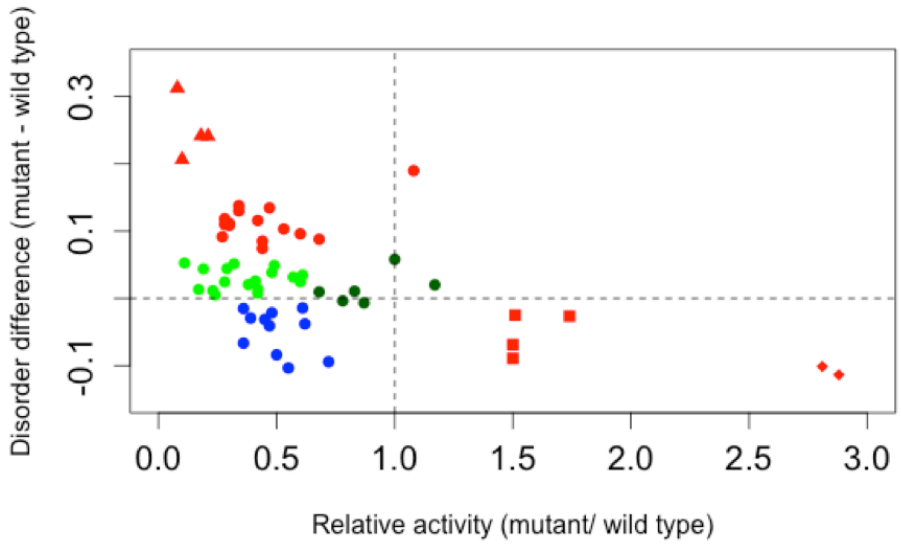


Figure 4.2: Relation between relative experimental activity A_{rel} and disorder difference as a function of the relative activity using IUPred_long

4.1.2 Stability in τ_1 core

We studied the unfolding of the peptide fragments of the wild type and mutants of the τ_1 core of the Glucocorticoid Receptor. We used CHARMM program with CHARMM36 force field, described in the previous chapter in 3.3.

In figure 4.3, it is the represented a typical dependence between the mean helicity of the α-helix of ten runs and time, for two temperatures. We set up the runs up to 100ns. The helicity was measured by Kabsch and Sander's definition of α-helix. If the helicity is near 1, the peptide is close to being in α-helix conformation. On the opposite, if the helicity is close to 0, the peptide is mostly unfolded. Since we simulated the unfolding of the peptide, it was expected a decrease of helicity with the increase of time. Comparing the different temperatures, the data from the simulations at 400K point to a faster unfolding, which it is expected.

We consider a peptide to be more stable, if it shows a slower decrease of helicity and vice-versa. Here, it will be introduced a new concept which is helicity first passage time t_h. This parameter represents the time (in ns) the helicity takes to decreases to 0.5, in a number of computational experiments (run). Figure 4.4 is a typical histogram with the frequency of t_h for two temperatures. In agreement with the helicity plot (in figure 4.3), data from the simulations at 400K have in general an lower t_h values (in figure 4.3) comparing to the simulations at 360K.

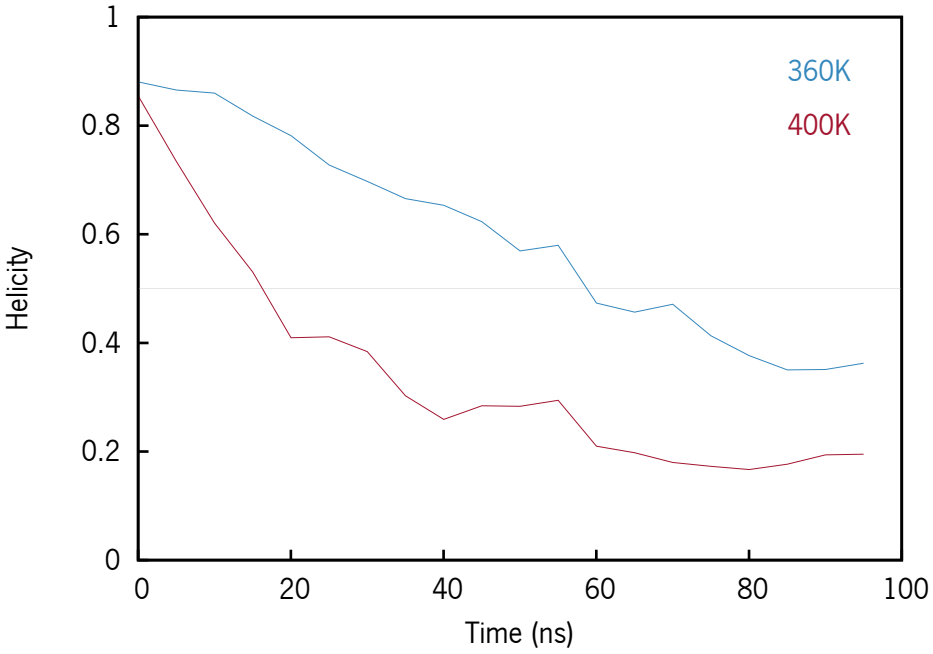


Figure 4.3: Helicity of τ_1 core of Glucocorticoid Receptor (wild type) during 100ns. Each point is an average of 5ns, in order to smooth the curve, and an average of 10runs. The data represented is simulations at 360K (blue) and 400K (red).

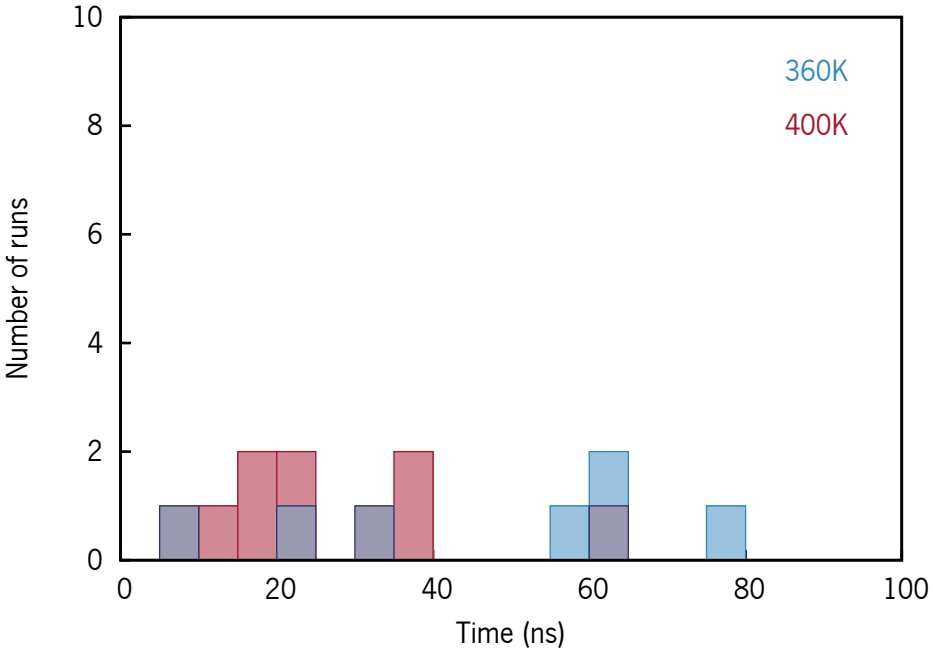


Figure 4.4: Frequency of the helicity first passage time t_h of τ_1 core of Glucocorticoid Receptor (wild type) during 100ns. The data represented are the simulations at 360K (blue) and 400K (red).

Table 4.1: The wild type and mutants and their respective activity, mean of helicity first passage time and the number of runs used to calculate the first passage time.

Mutants	A_{rel}	$\langle t_h \rangle$ (ns) at 360K	$\langle t_h \rangle$ (ns) at 400K	N at 360K	N at 400K
WT	100	43.58 ± 3.64	24.01 ± 1.60	7	10
L197P	34	18.01 ± 1.74	9.01 ± 0.99	10	10
F199E	34	54.30 ± 4.66	27.01 ± 1.60	7	10
F191E	29	50.01 ± 3.40	35.51 ± 2.03	7	10
L194P	28	37.51 ± 4.46	17.01 ± 1.57	6	10
L197E	30	51.68 ± 4.73	35.72 ± 4.66	6	7
F191D	28	46.26 ± 7.63	12.01 ± 0.67	4	10
I193D	27	45.84 ± 4.13	27.01 ± 3.01	6	10
F191A	44	43.76 ± 3.69	19.01 ± 1.66	8	10
D196Y	281	58.76 ± 5.63	45.01 ± 3.26	4	8
L197V	42	52.51 ± 5.60	25.01 ± 2.12	6	10
L194V	23	42.01 ± 5.08	29.01 ± 2.01	5	10
T190Y	150	21.68 ± 4.19	40.57 ± 2.60	3	9
T190F	150	52.15 ± 2.85	33.90 ± 2.20	7	9
I193F	151	62.51 ± 1.49	31.26 ± 2.52	8	8

Since each run takes its own time to reach half of the helicity, we calculated an average of helicity first passage time $\langle t_h \rangle$ by taking into account a maximum of 10 runs. We calculated $\langle t_h \rangle$ and the error associated $\sigma_{\langle t_h \rangle}$ according to (4.1) and (4.2), respectively. The number of runs taken into account is N. The table 4.1 shows the mutants, their activity, $\langle t_h \rangle$ for both temperature and number of runs used to calculate them.

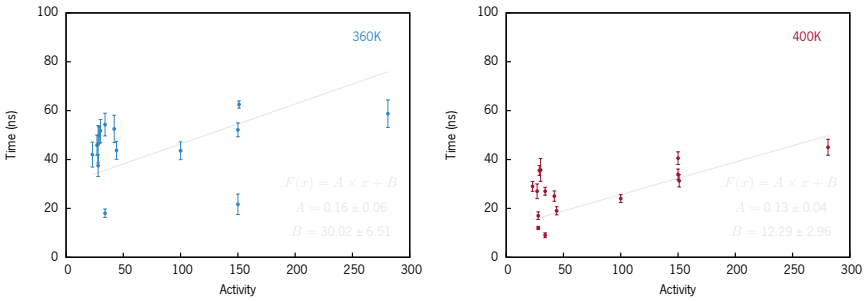
$$\langle t_h \rangle = \frac{\sum_{i=1}^{N} t_{hi}}{N} \tag{4.1}$$

$$\sigma_{\langle t_h \rangle} = \frac{\sqrt{\frac{\sum_{i=1}^{N} (t_{hi} - t_h)^2}{N-1}}}{\sqrt{N}} \tag{4.2}$$

From table 4.1, we plotted $\langle t_h \rangle$ as a function of the activity A_{rel} in figure 4.5. In figure 4.5a, the graph shows the data simulated at 360k. This appears to be relatively spread owing to two mutants with a distinguishably lower $\langle t_h \rangle$: the mutant L197P and the mutant T190Y.

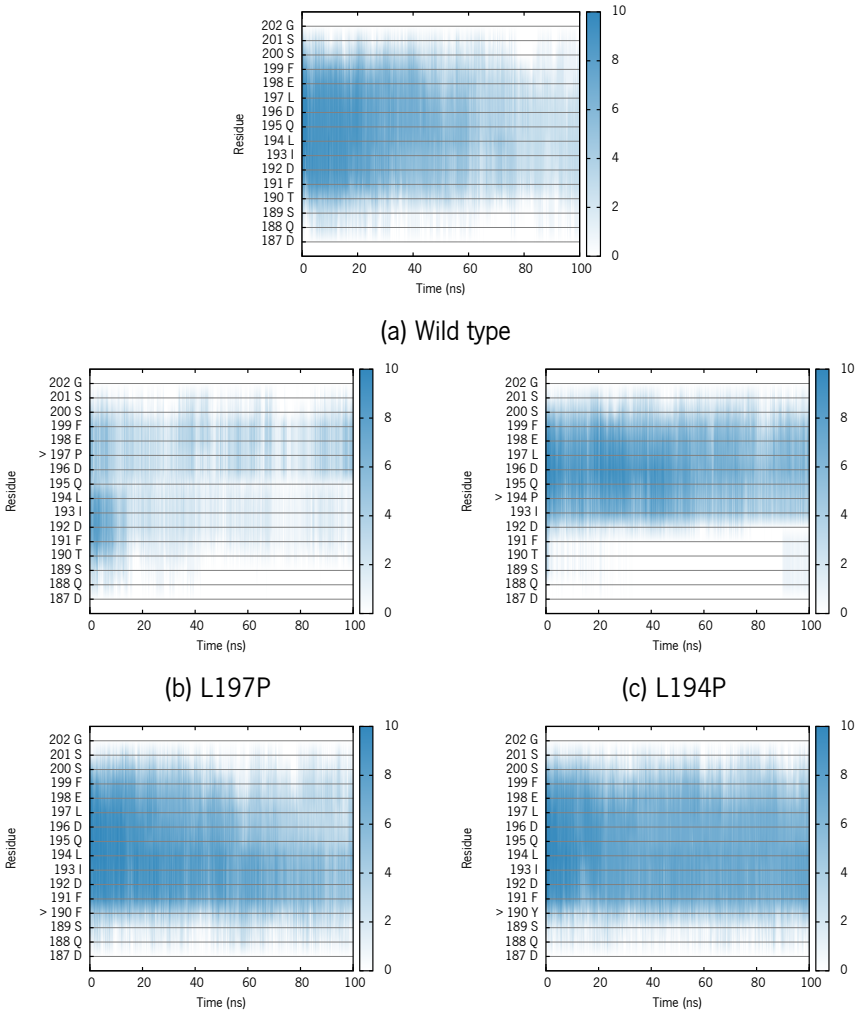
For a deeper understanding, we additionally plotted colourmaps in order to infer about the structure depending on the residues. In the colourmaps is represented the number of helical hydrogen bonds present in the 10 runs, by each residue in each time instant. If it is blue, the bond is present in most of the runs; in case of white, the bond is mostly absent. In the majority, the main helical region is situated in the centre of the peptide sequence and they behave similarly to the wild type. See

figure 4.6a.



(a) At 360K (b) At 400K

Figure 4.5: The $\langle t_h \rangle$ as a function of the activity A_{rel}. The simulations for $\langle t_h \rangle$ ran for 100ns. Each data point results of the average up to 10 runs.



(a) Wild type

(b) L197P (c) L194P

(d) T190F (e) T190Y

Figure 4.6: Sum of helical hydrogen bonds of each amino acid in every instant at 360K. The sum of helical hydrogen bonds corresponds to all helical hydrogen bonds of 10runs.

The main helical region of the mutant L197P (in figure 4.6b) looks divided in two and seems to be very unstable. Prolines do not have the backbone N-H group and this is because of the conformational restriction imposed by its side chain, which is a closed ring. The N-H group is involved in the hydrogen bonds and stabilises an α-helix. So this absence affects the conformation, which is clear in figures 4.6b. Through the colourmap of the mutant L197P becomes evident why its measurement of $\langle t_h \rangle$ is so low.

Curiously, by contrast with L197P (in figure 4.6b), the main helical region of the mutant L194P (in figure 4.6c) is more stable and shorter than the WT. Even though the main helical region of L194P is more stable, the remainder amino acids appear to be mostly unstructured. So, since the $\langle t_h \rangle$ accounts for the all region, L194P (in figure 4.6c) has a smaller $\langle t_h \rangle$ compared to WT (in figure 4.6a). With these two mutants, L197P and L194P, we can ratify the stabilisation effect of Prolines is not always the same.

Other interesting results are the mutations in T190 (in figures 4.6d and 4.6e). The mutants T190Y and T190F have high activity. According to our results, both also have a stable conformation and look to unfold from the c-end into the middle. This unfolding process is easily seen in figure 4.6d.

Concerning the mutant T190Y, the value obtained for $\langle t_h \rangle$ is definitely questionable. In the figure 4.6e, we can confirm the mutant T190Y has rather high stability. The reason for such a low measured $\langle t_h \rangle$ is it only takes three runs into account. We were just able to measure three times the t_h, presumably because the simulation time was not enough.

Since simulations at 400K point to a faster unfolding, generally we managed to measure the t_h in a higher number of runs per mutant. Consequently, $\langle t_h \rangle$ at 400K, in figure 4.5b, seem to be more accurate than the ones at 360k.

Every single mutation seems to have an individual effect of helicity/stability of the τ_1 core peptide fragment, the changes induced by mutations seem to be not only quantitive but also qualitative. That being so, it can be speculated that $\langle t_h \rangle$ cannot describe the effect of modification completely or inform about the helicity/order/stability of the τ_1 core region. It seems that a lot is unknown about structural-function relationship in Intrinsically Disordered regions like in τ_1 core of the Glucocorticoid Receptor.

4.2 c-Myc

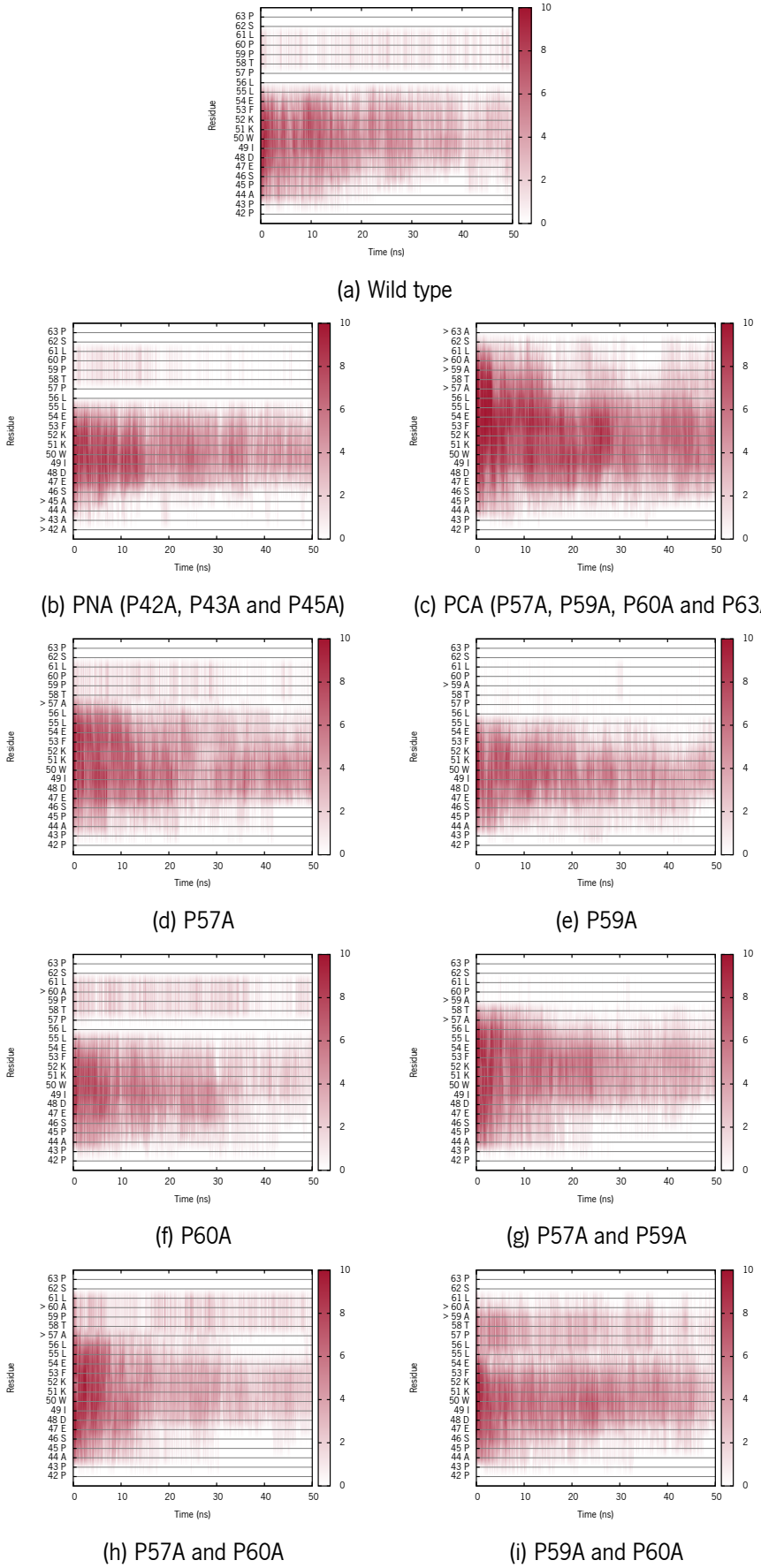
4.2.1 MBI alone

Before simulating MBI region of the c-Myc with a compound, we started by simulating MBI region alone. We adapted the protocol used with τ_1 core of Glucocorticoid Receptor to MBI of c-Myc, which is mentioned in the previous chapter in 3.3. We plotted the colourmaps used with τ_1 core of Glucocorticoid Receptor, but MBI region of c-Myc shows different results. The presence of a transient secondary structure within MBI was mentioned earlier in 1.2.2. In the colourmaps is represented the number of helical hydrogen bonds present in the 10 runs, by each residue in each time instant. If it is red, the bond is present in most of the runs; in case of white, the bond is mostly absent. From figure 4.7a, it is possible to perceive the two regions described as having helical properties and fluctuating extended character.

The Prolines contribution to the conformational stability of the helical transient secondary in Intrinsically Disordered proteins has been investigated. The replacement of a N-terminal flanking Proline is related to a decrease of α-helical content, while the replacement of a C-terminal flanking Proline is related to an increase α-helical content [49]. We simulated MBI fragment with a triple-mutation PNA (in figure 4.7b), in which Prolines in N-terminal flanking region were substituted with Alanines (P42A, P43A and P45A), and a quadruple-mutation PCA (in figure 4.7c), in which Prolines in C-terminal flanking region were substituted with Alanines (P57A, P59A, P60A and P63A). The obtained results were in agreement with the mention study on the Prolines [49], see figure 4.7b and 4.7c. In order to study more deeply the influence of Prolines in the structure, we simulated the mutants P57A, P59A and P60A and double mutations, P57A and P59A, P57A and P60A and P59A and P60A (in figure 4.7). Despite what it is described in the literature, we did not find a more detailed trend of the role of Prolines on the stabilisation of disordered regions of c-Myc.

4.2.2 MBI with the compound

The compound structure used here was suggested by Professor Roger Strömberg. We run simulation with MBI region of c-Myc peptide fragment close by the ligand. The ligand was design in order to favour the formation of the c-Myc-ligand complex. We monitored the contacts between the protein and the compound, to see if the ligand does indeed bind to the peptide as it was designed to do. The binding has two kinds of interactions: hydrogen bonds and hydrophobic contacts. Hence, we plotted the binding via these two interactions, in figure 4.9.



(a) Wild type

(b) PNA (P42A, P43A and P45A)

(c) PCA (P57A, P59A, P60A and P63A)

(d) P57A

(e) P59A

(f) P60A

(g) P57A and P59A

(h) P57A and P60A

(i) P59A and P60A

Figure 4.7: Sum of helical hydrogen bonds of each amino acid in every instant at 400K. The sum of helical hydrogen bonds corresponds to all helical hydrogen bonds of 10runs.

In graphs of figure 4.9, our data point to a stronger hydrophobic interaction consistent in all three temperatures. This temperature consistency is valid either for hydrogen bonds and for the hydrophobic contacts. Therefore, the analysis below is done for 330K, since the scale is expanded.

Looking from the ligand perspective, the range of occupancy levels is higher for hydrophobic contacts compared to hydrogen bonds. At 330K, the major hydrophobic contact occupancy is around 0.8 (in figure 4.9d), while it is around 0.3 for the highest hydrogen bond occupancy of the ligand (in figure 4.9c). Interestingly, the hydrophobic contact points are situated at backbone of compound while the hydrogen bonding are situated mostly at the extremities of the molecule. See figure 4.9e.

Looking at the contacts of MBI region of c-Myc (in figure 4.9a and 4.9b), we can verify again that the range of occupancy levels is higher for hydrophobic contacts. Out data suggest the ligand has the tendency to bind to the residues located at the middle of the MBI region, more specifically to the Isoleucine (I49), Tryptophan (W50) and Phenylalanine (F53), with an occupancy levels higher than 0.8. The hydrogen bonding is mostly done by Lysine (K52), Glutamic Acids (E54 and E47).

A possible top view of the interactions between the ligand and MBI peptide fragment is shown at figure 4.8. The results of this work point to understand if ligand can be a possible drug to regulate the funtion of c-Myc. Concerning the c-Myc results, they are still preliminary ones.

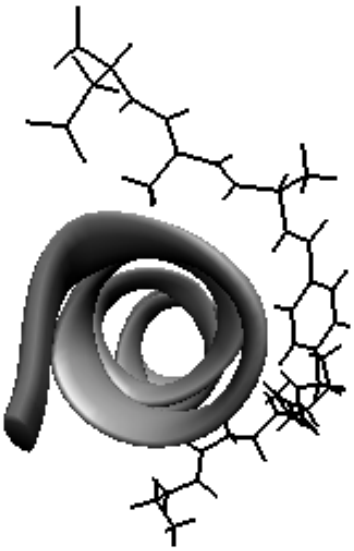
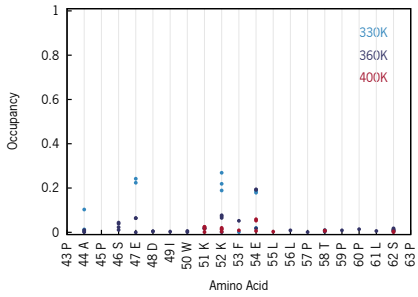
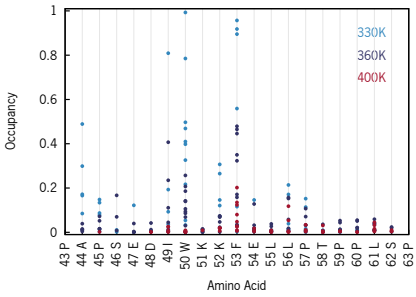


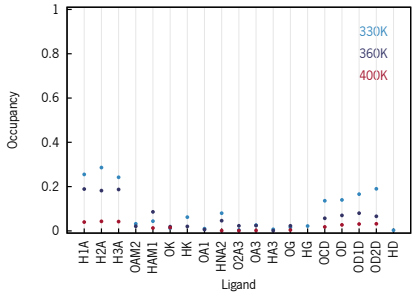
Figure 4.8: Schematic figure representing the top view the starting point of interaction between the c-Myc and ligand. The compound is supposedly linked in the middle of the helix as suggested by occupancy results



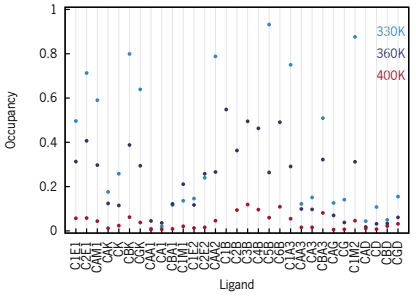
(a) Occupancy level for hydrogen bonds interaction of MBI region of c-Myc with the ligand



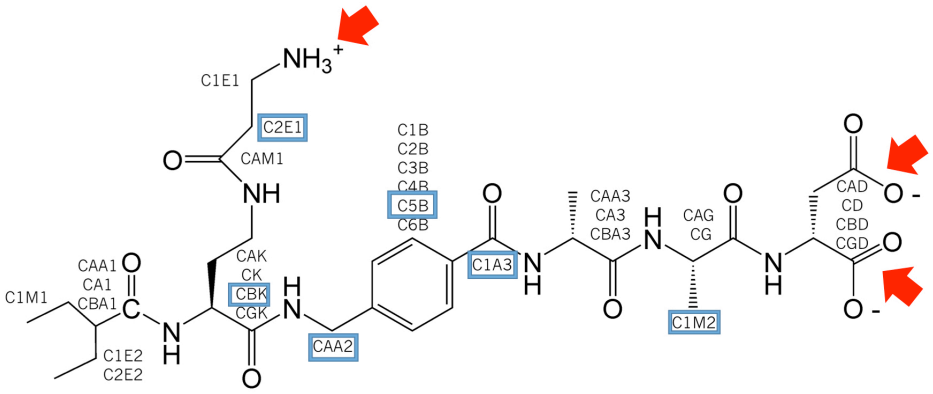
(b) Occupancy level for the hydrophobic contacts interaction of MBI region of c-Myc with the ligand



(c) Occupancy level for hydrogen bonds interaction of the ligand with MBI region of c-Myc



(d) Occupancy level for the hydrophobic contacts interaction of the ligand with MBI region of c-Myc



(e) Chemical structure of compound with red arrows representing hydrogen bonding and the hydrophobic points marked with a blue rectangle.

Figure 4.9: Representation of the interactions between MBI region of c-Myc and the ligand

5 Conclusion

This work intents to enlighten as to the conformation state of disordered regions of Glucocorticoid Receptor and c-Myc proteins.

The data on the τ_1 core of the Glucocorticoid Receptor showed an inverse correlation between relative activity and disorder. We found many single effects in the τ_1 core but not a general trend between the experimental biological relative activity and its stability. The data of disordered region MBI of c-Myc analysis showed Prolines have both a destabilising and stabilising effect. Here again we did not find a deeper trend on what is responsible for protein order/disorder of this c-Myc fragment. Research in Intrinsically Disordered proteins is a very complex task and lot of it is still unknown.

Analysis of the interactions between the c-Myc and the ligand showed that hydrophobic contacts are the most important kind of interaction, representing an occupancy levels higher than 0.7 at 330 K. Our data point to the fact that the ligand is bound to the residues located at the middle of the MBI fragment peptide more specifically to the Isoleucine (I49), Tryptophan (W50) and Phenylalanine (F53).

The results of this work set new directions to understand the potential of this ligand as a possible drug to regulate the funtion of c-Myc.

Bibliography

1. Li, J. et al. An Overview of Predictors for Intrinsically Disordered Proteins over 2010–2014. International journal of molecular sciences 16, 23446–23462 (2015).
2. He, B. et al. Predicting intrinsic disorder in proteins: an overview. Cell research 19, 929–949 (2009).
3. Chen, J. Towards the physical basis of how intrinsic disorder mediates protein function. Archives of biochemistry and biophysics 524, 123–131 (2012).
4. Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. Current opinion in structural biology 21, 432–440 (2011).
5. Uversky, V. N. Under-folded proteins: Conformational ensembles and their roles in protein folding, function, and pathogenesis. Biopolymers 99, 870–887 (2013).
6. Hansen, J. C., Lu, X., Ross, E. D. & Woody, R. W. Intrinsic protein disorder, amino acid composition, and histone terminal domains. Journal of Biological Chemistry 281, 1853–1856 (2006).
7. Campen, A. et al. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein and peptide letters 15, 956 (2008).
8. Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. Annual review of biochemistry 83, 553–584 (2014).
9. Hilser, V. J. & Thompson, E. B. Structural dynamics, intrinsic disorder, and allostery in nuclear receptors as transcription factors. Journal of Biological Chemistry 286, 39675–39682 (2011).
10. Varadi, M., Vranken, W., Guharoy, M. & Tompa, P. Computational approaches for inferring the functions of intrinsically disordered proteins. Frontiers in molecular biosciences 2 (2015).
11. Stanley, N., Esteban-Martín, S. & De Fabritiis, G. Progress in studying intrinsically disordered proteins with atomistic simulations. Progress in biophysics and molecular biology 119, 47–52 (2015).

12. Rawat, N. & Biswas, P. Hydrogen bond dynamics in intrinsically disordered proteins. The Journal of Physical Chemistry B 118, 3018–3025 (2014).

13. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. Nature reviews Molecular cell biology 6, 197–208 (2005).

14. Gibbs, E. B. & Showalter, S. A. Quantitative biophysical characterization of intrinsically disordered proteins. Biochemistry 54, 1314–1326 (2015).

15. Kim, S. et al. Probing allostery through DNA. Science 339, 816–819 (2013).

16. Cheng, Y. et al. Rational drug design via intrinsically disordered protein. Trends in biotechnology 24, 435–442 (2006).

17. Paul, S. M. et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature reviews Drug discovery 9, 203–214 (2010).

18. Nilsson, J., Grahn, M. & Wright, A. P. Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. Genome Biol 12, R65 (2011).

19. Herdegen, T & Leah, J. Inducible and constitutive transcription factors in the mammalian nervous system: control of gene expression by Jun, Fos and Krox, and CREB/ATF proteins. Brain Research Reviews 28, 370–490 (1998).

20. Jolma, A. Determination of transcription factor binding specificities (Inst för biovetenskaper och näringslära/Dept of Biosciences and Nutrition, 2015).

21. Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. Nature Reviews Genetics 13, 613–626 (2012).

22. Lammens, T., Li, J., Leone, G. & De Veylder, L. Atypical E2Fs: new players in the E2F transcription factor family. Trends in cell biology 19, 111–118 (2009).

23. Almlöf, T., Gustafsson, J.-A. & Wright, A. Role of hydrophobic amino acid clusters in the transactivation activity of the human glucocorticoid receptor. Molecular and cellular biology 17, 934–945 (1997).

24. Almlöf, T., Wallberg, A. E., Gustafsson, J.-Å. & Wright, A. P. Role of important hydrophobic amino acids in the interaction between the glucocorticoid receptor τ1-core activation domain and target factors. Biochemistry 37, 9586–9594 (1998).

25. Dahlman-Wright, K. et al. Structural characterization of a minimal functional transactivation domain from the human glucocorticoid receptor. Proceedings of the National Academy of Sciences 92, 1699–1703 (1995).

26. Castro-Vale, I., van Rossum, E. F., Machado, J. C., Mota-Cardoso, R. & Carvalho, D. Genetics of glucocorticoid regulation and posttraumatic stress disorder—What do we know? Neuroscience & Biobehavioral Reviews 63, 143–157 (2016).

27. Helsen, C. & Claessens, F. Looking at nuclear receptors from a new angle. Molecular and cellular endocrinology 382, 97–106 (2014).

28. John, K., Marino, J. S., Sanchez, E. R. & Hinds, T. D. The glucocorticoid receptor: cause of or cure for obesity? American Journal of Physiology-Endocrinology and Metabolism 310, E249–E257 (2016).

29. Wärnmark, A., Treuter, E., Wright, A. P. & Gustafsson, J.-A. Activation functions 1 and 2 of nuclear receptors: molecular strategies for transcriptional activation. Molecular endocrinology 17, 1901–1909 (2003).

30. McEwan, I. J., Dahlman-Wright, K., Ford, J. & Wright, A. P. Functional interaction of the c-Myc transactivation domain with the TATA binding protein: evidence for an induced fit model of transactivation domain folding. Biochemistry 35, 9584–9593 (1996).

31. Andresen, C. et al. Transient structure and dynamics in the disordered c-Myc transactivation domain affect Bin1 binding. Nucleic acids research 40, 6353–6366 (2012).

32. Tu, W. B. et al. Myc and its interactors take shape. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 1849, 469–483 (2015).

33. Smith, S. M., Anastasi, J., Cohen, K. S. & Godley, L. A. The impact of MYC expression in lymphoma biology: beyond Burkitt lymphoma. Blood Cells, molecules, and diseases 45, 317–323 (2010).

34. Helander, S. et al. Pre-Anchoring of Pin1 to Unphosphorylated c-Myc in a Fuzzy Complex Regulates c-Myc Activity. Structure 23, 2267–2279 (2015).

35. Nair, S. K. & Burley, S. K. X-ray structures of Myc-Max and Mad-Max recognizing DNA: molecular bases of regulation by proto-oncogenic transcription factors. Cell 112, 193–205 (2003).

36. Feynman, R., Leighton, R. & Sands, M. The Feynman Lectures on Physics, Vol. I: The New Millennium Edition: Mainly Mechanics, Radiation, and Heat vol. 1. isbn: 9780465040858 (Basic Books, 2015).

37. Karplus, M., Petsko, G. A., et al. Molecular dynamics simulations in biology. Nature 347, 631–639 (1990).

38. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. Nature Structural & Molecular Biology 9, 646–652 (2002).

39. Burendahl, S. Molecular Dynamic Studies of Nuclear Receptors Ligand Binding Domain (Biovetenskaper och näringslära/Biosciences and Nutrition, 2009).

40. Hart, K. Molecular dynamics of protein-nucleic acid complexes (Biovetenskaper och näringslära/Biosciences and Nutrition, 2010).

41. Allen, M. & Tildesley, D. Computer Simulation of Liquids isbn: 9780198556459 (Clarendon Press, 1989).

42. Brooks, B. R. et al. CHARMM: the biomolecular simulation program. Journal of computational chemistry 30, 1545–1614 (2009).

43. Gould, H., Tobochnik, J. & Christian, W. An Introduction to Computer Simulation Methods: Applications to Physical Systems isbn: 9780805377583 (Pearson Addison Wesley, 2007).

44. PONDR. 2007. <http://www.pondr.com/pondr-tut2.html>.

45. Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K. & Uversky, V. N. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1804, 996–1010 (2010).

46. IUPred. <http://iupred.enzim.hu>.

47. IUPred. <http://iupred.enzim.hu/Help.php>.

48. ESPRITZ. 2011. <http://147.162.170.247:8080/espritz/help_pages/help.html>.

49. Lee, C. et al. Contribution of proline to the pre-structuring tendency of transient helical secondary structure elements in intrinsically disordered proteins. Biochimica et Biophysica Acta (BBA)-General Subjects 1840, 993–1003 (2014).